Educational Measurement: Issues and Practice
Fall 2011, Vol. 30, No. 3, pp. 29–40

An NCME Instructional Module on Subscores

Sandip Sinharay, Gautam Puhan, and Shelby J. Haberman, *Educational Testing Service*

The purpose of this ITEMS module is to provide an introduction to subscores. First, examples of subscores from an operational test are provided. Then, a review of methods that can be used to examine if subscores have adequate psychometric quality is provided. It is demonstrated, using results from operational and simulated data, that subscores have to be based on a sufficient number of items and have to be sufficiently distinct from each other to have adequate psychometric quality. It is also demonstrated that several operationally reported subscores do not have adequate psychometric quality. Recommendations are made for those interested in reporting subscores for educational tests.

Keywords: augmented subscore, classical test theory, diagnostic score, item response theory, mean squared error

What Are Subscores and Why Should Anyone Care About Them?

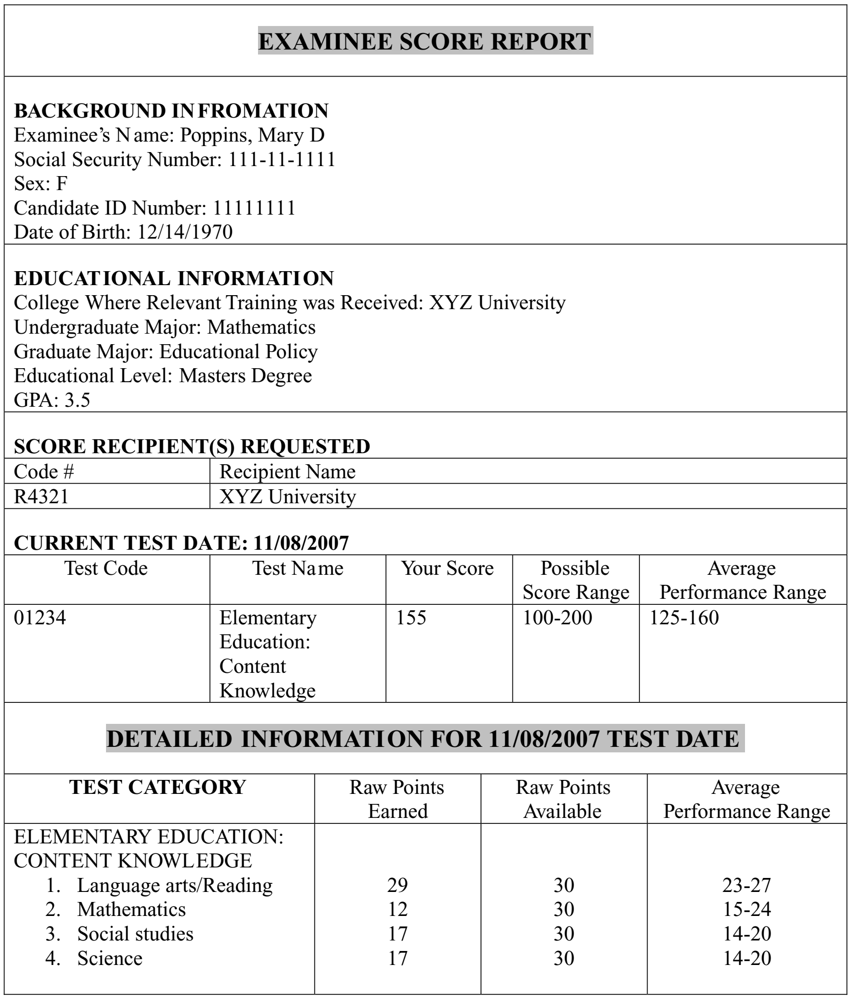
Figure 1 shows a hypothetical score report of an imaginary examinee (Mary D. Poppins) on the Elementary Education: Content Knowledge test that is designed for prospective teachers of children in primary through upper elementary school grades (Educational Testing Service, 2008). This test contains only multiple-choice (MC) items and was considered in Sinharay (2010) and Puhan, Sinharay, Haberman, and Larkin (2010). There are four subareas/categories covered in the test: language arts/reading, mathematics, social studies, and science. The score report shows the total (scaled) score earned by Mary. The bottom portion of the score report shows the raw points (or scores) earned by Mary on the different categories of the test. The figure also shows the raw points available (or, the total number of items, as this is a MC test) in each category, which is 30 for all categories, and the range of scores obtained by the middle 50% of the population of examinees who took the test form within a certain time-period ("average performance range"). The bottom portion of the score report represents a typical subscore report for examinees—the scores for the categories are the subscores. For example, Mary obtained a subscore of 12 on mathematics and a subscore of 17 on science. The common perception is that (a) subscores provide trustworthy information about the examinee's strengths and weaknesses, and (b) the examinee will work harder on the categories on which she performed poorly and hence improve in those areas. For example, Mary scored only 12 out of a maximum possible 30 on mathematics and it is expected that she would work harder on the content of this area and improve. See Figures 17–21 of Goodman and Hambleton (2004) for examples of several operational score reports that include subscores.

The nine panels of Figure 2 show the subscores of nine examinees on one recent form of the Elementary Education: Content Knowledge test. We will use these data on several occasions in this module. The total scores of these examinees, shown in the titles of the panels, are the 10th, 20th, . . ., and 90th percentiles of the distribution of the total scores for the form. The subscores are shown using vertical bars (explanation of the solid squares in the panels will be provided later). Note that the subscores shown for the examinee with total score = 75 in Figure 2 are identical to those of Mary D. Poppins. The values of the average performance range in Figure 1 indicate that the language arts/reading subscore is larger than the other subscores on an average. The means of the subscores are 25.0, 19.4, 17.0, and 17.2, respectively. Figure 2 shows several interesting patterns. For example, the examinee whose total score is 62 performed considerably better in social studies than in mathematics while the pattern is the opposite for the examinee with a total score of 84 and there is another examinee, one with total score of 91, who performed comparatively poorly in mathematics like Mary. Although the subscores in Figure 1 represent four different subject areas, it is possible to define subscore as scores on subsections within a single subject area (e.g., algebra and geometry subsections on a mathematics test) or in any other way the test developers think is appropriate. Wainer, Sheehan, and Wang (2000) gave examples of how the same test can be used to report different sets of subscores. For example, for the same test, they considered a set of seven subscores based on content classification (science, fine arts, mathematics, reading/language arts, sociology, and physical education) and another set of ten subscores based on the primary skill area addressed by the items (content knowledge, solving social/emotional/affective and classroom management problems, selecting alternative instructional strategies, etc.).

Subscores are of increasing interest in educational testing due to their potential remedial and instructional benefits. According to the National Research Council report "Knowing What Students Know" (2001), the target of assessment is to provide particular information about an examinee's

Sandip Sinharay, Gautam Puhan, and Shelby J. Haberman, Educational Testing Service, Princeton, NJ 08541; ssinharay@ets.org.

Any opinions expressed in this article are those of the authors and not necessarily of Educational Testing Service.



EXAMINEE SCORE REPORT

BACKGROUND INFROMATION
Examinee's Name: Poppins, Mary D
Social Security Number: 111-11-1111
Sex: F
Candidate ID Number: 11111111
Date of Birth: 12/14/1970

EDUCATIONAL INFORMATION
College Where Relevant Training was Received: XYZ University
Undergraduate Major: Mathematics
Graduate Major: Educational Policy
Educational Level: Masters Degree
GPA: 3.5

SCORE RECIPIENT(S) REQUESTED

Code #	Recipient Name
R4321	XYZ University

CURRENT TEST DATE: 11/08/2007

Test Code	Test Name	Your Score	Possible Score Range	Average Performance Range
01234	Elementary Education: Content Knowledge	155	100-200	125-160

DETAILED INFORMATION FOR 11/08/2007 TEST DATE

TEST CATEGORY	Raw Points Earned	Raw Points Available	Average Performance Range
ELEMENTARY EDUCATION: CONTENT KNOWLEDGE			
1. Language arts/Reading	29	30	23-27
2. Mathematics	12	30	15-24
3. Social studies	17	30	14-20
4. Science	17	30	14-20

FIGURE 1. A hypothetical score report for an examinee showing performance on the total test and the different test categories.

knowledge, skill, and abilities. Subscores have the potential to provide such information. The U.S. Government's No Child Left Behind (NCLB) Act of 2001 demands, among other things, that students should receive diagnostic reports that allow teachers to address their specific academic needs; subscores could be used in such a diagnostic report. Naturally, there is substantial pressure on testing programs to report subscores, both at the individual examinee level and at aggregate levels such as at the level of institutions or states. Subscores are reported by several large-scale testing programs, such as SAT®, ACT®, Praxis, and LSAT.

The purpose of this module is to review the research, most of which is recent, on subscores. A brief review of several methods that can be used to examine the psychometric quality of subscores is provided, followed by a closer examination of some existing subscores and their usefulness. We then provide a brief discussion of alternatives to reporting subscores. Finally, some recommendations are made for researchers and practitioners interested in the issue of subscore reporting.

Examining Whether Subscores Have Adequate Psychometric Quality

Despite the demand for and apparent usefulness of subscores, they have to satisfy certain quality standards in order for them to be reported. Standard 5.12 of the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999) states that "Scores should not be reported for individuals unless the validity, comparability, and reliability of such scores have been established." The standard applies to subscores as well. Further, Standard 1.12 of the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) states that if a test provides more than one score, the distinctiveness of the separate scores should be demonstrated. These quality standards are not too demanding if one considers that just as inaccurate information at the total test score level may lead to decisions with damaging consequences, inaccurate information at the subscore level can also lead to incorrect remediation decisions resulting in large and needless expenses for examinees, states, or institutions.

Why the Subscores from Educational Tests Often Lack Psychometric Quality

It is not uncommon to observe decent reliabilities of subscores on personality inventories designed to measure specific personality traits, such as anxiety, hostility, trust, etc. For example, Goldberg (1999) reported that for the revised NEO Personality Inventory, the reliabilities of the 30 subscores, each of which consisted of eight items, ranged

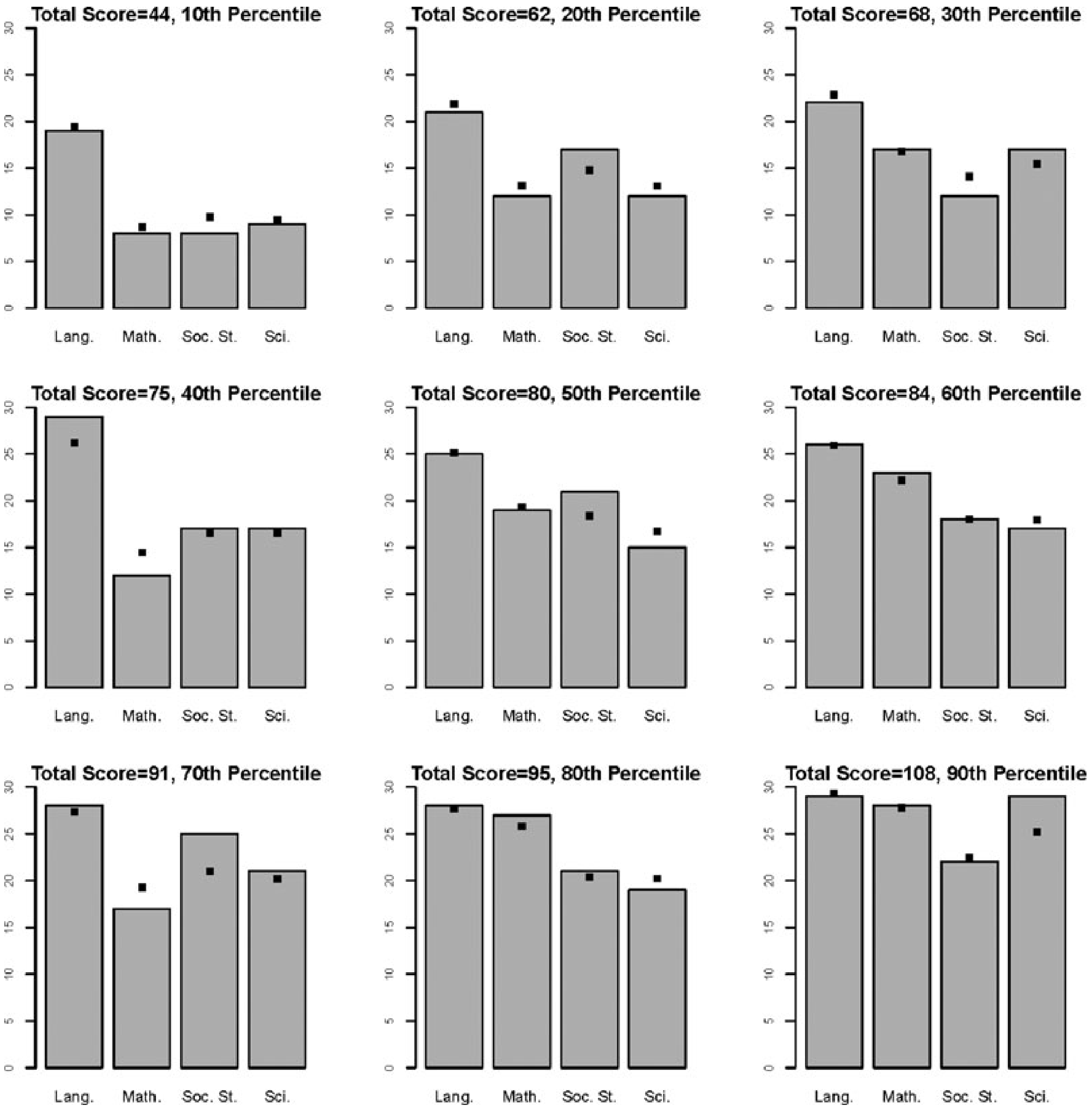


FIGURE 2. The subscores (vertical bars) and weighted averages (solid squares) of nine examinees from one form of the Elementary Education: Content Knowledge test.

from .61 to .85 and the mean reliability was .75. Considering the relatively small number of items in each of the subscales, these reliabilities seem reasonably high for diagnostic use. Personality inventories are often designed to measure fewer but much more focused traits, such as aggression and gregariousness, and are usually reliable. The subscores found in educational measurement often consist of only a few items. For example, the state test score report shown in Figure 20 of Goodman and Hambleton (2004) includes seven subscores that are based on only five items each. Such scores are most often outcomes of retrofitting, which refers to reporting of subscores from tests that were designed to measure only one overall skill. These scores are usually provided to comply with clients' requests for more diagnostic information on examinees. Because these tests have been constructed specifically to measure a single construct, little reason exists to expect useful subscores. In addition, practitioners often ignore the fact that a subscore in educational measurement refers to a domain area that is usually much broader than those covered by subscores in personality inventories. Consider a certification test for prospective teachers of elementary education that measures content knowledge in domain areas such as mathematics, science, and reading. It is often observed that, within each domain area, all items do not measure a single ability. For example, the mathematics domain may have numerous smaller sub-domains such as problem solving, algebra, geometry, data organization using pie charts, bar graphs, etc. Therefore a substantial number of items will be typically required in each sub-domain to make the mathematics subscore reliable.

Sinharay and Haberman (2008b) discussed that data may not provide information as fine-grained as suggested by cognitive theory or as hoped for by the testing practitioner. A theory of response processes based on cognitive psychology may suggest several skills, but a test includes a limited number of items and the test may not have enough items to provide adequate information about all of these skills. For example, for the iSkills^{TM} test (e.g., Katz, Attali, Rijmen, & Williamson, 2008), an expert committee identified seven performance areas that they thought comprised Information and Communications Technology literacy skill. However, a factor analysis of the data revealed only one factor and confirmatory factor models in which the factors corresponded to performance areas or a combination thereof did not fit the data (Katz et al., 2008). As a result, only an overall Information and Communications Technology literacy score is reported for the test. Clearly, an investigator attempting to report subscores has to make an informed judgment on how much evidence the data can reliably provide and report only as much information as is reliably supported.

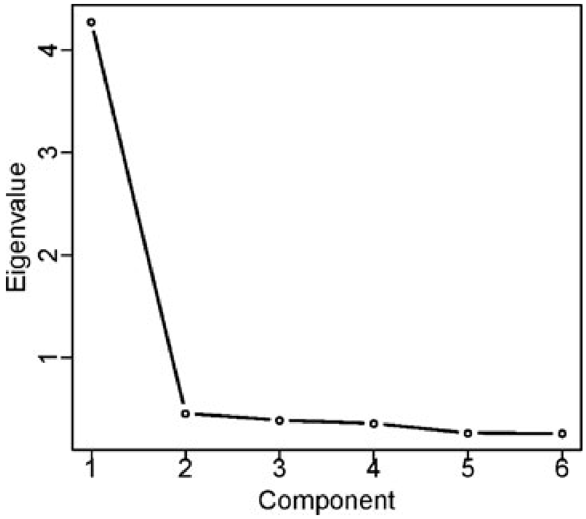


FIGURE 3. A scree plot of the eigenvalues computed from the correlations between the six subscores of a basic skills test.

Methods

Researchers have suggested several methods for examining if subscores have adequate psychometric quality—those methods are discussed next.

Application of Factor Analysis

Several researchers, such as Stone, Ye, Zhu, & Lane (2010), Wainer et al. (2001), and Sinharay, Haberman, and Puhan (2007) employed factor analysis to determine whether subscores are distinct enough to be reported. The purpose of factor analysis is to discover simple patterns in the pattern of relationships among several observed variables. In particular, it seeks to discover if the observed variables can be explained largely or entirely in terms of a small number of variables called factors. A simple factor-analytic approach to evaluate whether the subscores are distinct enough would be to compute the eigenvalues from the correlation matrix of the subscores (or from the correlation matrix of the items). If most of the eigenvalues computed from the correlation matrix of the subscores are smaller than 1 or if a scree plot of these eigenvalues shows that the eigenvalues abruptly levels out at some point, then the number of factors in the data is less than the number of subscores and the claim of several distinct subscores is probably not justified. The presence of multiple factors, that is, multiple large eigenvalues, would support the reporting of subscores. For example, Sinharay et al. (2007) computed the eigenvalues from the 6 × 6 correlation matrix of six reported subscores from two forms of a basic skills test. They found that the largest eigenvalue was 4.3 for both forms while the remaining five eigenvalues were smaller than .5, suggesting that the test is essentially unidimensional and the claim of 6 distinct subscores is probably not justified (see Figure 3 for a scree plot of the six eigenvalues from one form). Similarly, Stone et al. (2010) reported, using an exploratory factor analysis method on the inter-item correlation matrix, the presence of only one factor in the Spring 2006 assessment of the Delaware State Testing Program 8th grade mathematics assessment—so subscores should not be reported for the assessment. Use of factor analysis involves several issues such as determining whether to perform the factor analysis at the item level or at the item parcel level or the subscore level, determining whether to use exploratory or confirmatory factor analysis, and determining which tools to use to determine the number of factors.

Application of the Beta-Binomial Model

Harris and Hanson (1991) suggested the method of fitting a mathematical model named the beta-binomial model (Lord, 1965) to the observed subscore distributions to determine if the subscores have added value over and beyond the total score. Consider a test with two subscores. If the bivariate distribution of the two subscores computed under the assumption that the corresponding true subscores are functionally related provides an adequate fit to the observed bivariate distribution of subscores, the true subscores are functionally related and therefore do not provide any added value. Harris and Hanson used a chi-square-type statistic in their data example to determine the goodness of fit of the bivariate distribution of the two subscores under the assumption that the corresponding true subscores are functionally related to the observed bivariate distribution of subscores.

Fitting of Multidimensional Item Response Theory Models

Another way to examine if subscores have added value is to fit a multidimensional item response theory (MIRT) model (e.g., Reckase, 1997; Ackerman, Gierl, & Walker, 2003) to the data. MIRT is a tool to model examinee responses to a test that measures more than one ability, for example, a test that measures both mathematical and verbal ability. In MIRT, the probability of a correct item response is a function of several abilities, rather than a single measure of ability. To examine if subscores have added value, one can perform a statistical test of whether a MIRT model provides a better fit to the data than a unidimensional IRT model. See von Davier (2008) for a demonstration of this sort of use of MIRT.

Application of DIMTEST and DETECT

Researchers such as Ackerman and Shu (2009) used dimensionality assessment software programs such as DIMTEST

(Stout, 1987) and DETECT (Zhang & Stout, 1999) to determine the usefulness of subscores. DETECT uses an algorithm that searches through all of the possible item cluster partitions to find the one that maximizes the DETECT statistic. Based on results from a simulation study, Kim (1994) provided guidelines to interpret the DETECT statistic. According to these guidelines, if the DETECT statistic is less than .10, then the data can be considered as unidimensional. Values between .10 and .50 would indicate a weak amount of dimensionality, values between .51 and 1.00 would indicate a moderate amount of dimensionality, and values higher than 1.00 would indicate a strong amount of dimensionality. DIMTEST implements a hypothesis testing procedure to evaluate the lack of unidimensionality in data from a test. It assesses the statistical significance of the possible dimensional distinctiveness between two specified subtests (the Assessment Subtest or AT and the Partitioning Subtest or PT). The test statistic T calculated by DIMTEST represents the degree of dimensional distinctiveness between these two specified subsets. For example, if the testing practitioner wants to test if the math items on a general ability test are dimensionally distinct from the rest of the items in the test, then the math items can form the assessment subtest and the remaining items can form the partitioning subtest and if the DIMTEST index T is significant, it would indicate that the two subsets are dimensionally distinct.

Application of the Classical-Test-Theory-Based Method of Haberman

Haberman (2008a), taking a classical test theory (CTT) viewpoint, assumed that a reported subscore is intended to be an estimate of the true subscore s_t and considered the following estimates of the true subscore:

- An estimate $s_s = \bar{s} + \alpha(s - \bar{s})$ based on the observed subscore s, where $\bar{s}$ is the average subscore for the sample of examinees and α is the reliability of the subscore.
- An estimate $s_x = \bar{s} + c\,(x - \bar{x})$ based on the observed total score x, where $\bar{x}$ is the average total score and c is a constant that depends on the reliabilities and standard deviations of the subscore and the total score and the correlations between the subscores.

Let us consider the mathematics subscore from the form of the Elementary Education: Content Knowledge test considered earlier. The mean and reliability of the subscore were 19.4 and .85, respectively. The mean and reliability of the total test score were 78.6 and .91, respectively. Further, the standard deviation of the mathematics subscore and the total test were 3.4 and 14.7, respectively and the constant $c = .31$ was calculated using formulas given in Haberman (2008a) and in the self-test later. Thus, for Mary D. Poppins, who obtained a mathematics subscore of 12 (i.e., 2.2 standard deviation less than the subscore mean) and a total score of 75 (i.e., only .2 standard deviation less than the total score mean), the estimate of the true subscore based on her observed subscore would be $19.4 + .85\,(12 - 19.4) = 13.1$, and the estimate of the true subscore based on her observed total score would be $19.4 + .31\,(75 - 78.6) = 18.3$, respectively.

Because Mary's mathematics subscore is unexpectedly low given her observed total score, an estimate of her true mathematics subscore based on her observed mathematics subscore is much lower than an estimate based on her observed total score.

To determine whether subscores have added value over the total score, Haberman (2008a) suggested the use of the proportional reduction in mean squared error or PRMSE. Computational details about the method can be found in Haberman. The PRMSESs lie between 0 and 1. The larger the PRMSE, the more accurate is the corresponding estimate. A larger PRMSE is equivalent to a smaller mean squared error in estimating the true subscore and hence is desirable. We denote the PRMSE for s_s and s_x as PRMSE$_s$ and PRMSE$_x$, respectively. The quantity *PRMSE$_s$* can be shown to be exactly equal to the reliability of the subscore. Haberman recommended that the subscore "provides added value over the total score" if and only if PRMSE$_s$ is larger than PRMSE$_x$; that is, the subscore "provides added value over the total score" if and only if the observed subscore performs better than the observed total score in terms of mean squared error in estimating the true subscore. Mathematically, PRMSE$_s$ is larger than PRMSE$_x$if and only if the correlation between the true subscore and the observed subscore is larger than the correlation between the true subscore and observed total score. Sinharay et al. (2007) discussed why this strategy suggested by Haberman is reasonable and how it ensures that a subscore satisfies professional standards. Conceptually, if the subscore is highly correlated with the total score, i.e., the subscore and the total score measure the same basic underlying skill(s), then the subscore does not provide any added value over what is already provided by the total score. A subscore is more likely to have added value if it has high reliability and it is distinct from the other subscores (Haberman, 2008a; Sinharay, 2010). For the aforementioned form from the Elementary Education: Content Knowledge test, the values of PRMSE$_s$ or subscore reliability are .71, .85, .67, and .72, respectively, while the values of PRMSE$_x$ are .72, .76, .74, and .79, respectively. Thus only the mathematics subscore of the test has added value.

The computations involved in the method of Haberman (2008a) are simple and involve only the sample variances, correlations, and reliabilities of the total score and the subscores. Some details of the computations are provided in two questions of the self-test. The computations of the PRMSEs involve the disattenuated correlations among the subscores, where the disattenuated correlation between two subscores is equal to the simple correlation between them divided by the product of the square roots of the reliabilities of the two subscores. For some data sets, it is possible to have disattenuated correlations between subscores larger than 1, typically because one or both reliabilities are low, and, as a result, PRMSEs larger than 1. In these cases, it is concluded that the subscores do not have any added value over the total score.

How Often Do Subscores Have Adequate Psychometric Quality?

Literature Review

Sinharay (2010) performed an extensive survey regarding whether subscores or section scores have added value over the total score for data from 25 operational tests, where the value added by a subscore was determined by the method of Haberman (2008a). Of the 25 tests, 16 had no subscores with added value even though several of them report subscores operationally. Even among the remaining nine tests, all of which had at least an average of 24 items contributing to each subscore, only some of the subscores had added value. Sinharay also performed a detailed simulation study to find out when subscores can be expected to have added value. The simulation study showed that in order to have added value, subscores have to (a) be based on a sufficient number of (roughly 20) items, and (b) be sufficiently distinct—the disattenuated correlation between subscores has to be less than about .85.

Stone et al. (2010) reported, using an exploratory factor analysis method, the presence of only one factor in the Spring 2006 assessment of the DSTP 8th grade mathematics assessment. Harris and Hanson (1991), using their aforementioned method of fitting beta-binomial distributions to the observed subscore distributions, found subscores to have little added value for the English and mathematics tests from the P-ACT+ examination. Wainer et al. (2001) performed factor analysis and an examination of the reliability of the subscores on data from one administration of the Just-in-Time Examination conducted by the American Production and Inventory Control Society (APICS) certification. They concluded that the six subscales in the APICS examination did not appear to measure different dimensions of individual differences. However, they found a tryout form of the North Carolina Test of Computer Skills that has four subscales to be not as unidimensional as the APICS examination and an application of the method of Haberman (2008a) to the data reveals that three of the four subscores have added value over the total score. Wainer et al. (2000) considered the problem of constructing skills-based subscores for the Education in the Elementary School Assessment that is designed for prospective teachers of children in primary grades (K-3) or upper-elementary/middle-school grades (4–8). They concluded, mostly from an analysis of reliability of the subscores, that the "test's items were fiercely unidimensional, and so any set of subscales that were chosen would yield essentially the same information." Ackerman and Shu (2009), using DIMTEST and DETECT, found subscores not to be useful for a fifth grade end-of-grade assessment. So, based on these studies, it would appear that subscores on operational tests have more often been found not to be useful than to be useful.

There is a lack of studies that demonstrated the validity of inferences made from the subscores. For example, there is little evidence showing that subscores are related to other external criteria. The only exception is Wallmark (unpublished data, 1981), who showed that GRE Psychology subscores are more strongly related to the number of courses taken in relevant subjects rather than in other areas. There is also a lack of evidence regarding the usefulness of the feedback from subscores in improving the future performance of the examinees. Haberman (2008b) demonstrated via theoretical derivations that the incremental validity of subscores is limited when subscores are either not reliable or are highly correlated with the total scores. More research is needed on this area.

Demonstration Using a Data Set

To demonstrate the problems with subscores consisting of only a few items, let us consider data from the aforementioned form of the Elementary Education: Content Knowledge test. We ranked the questions on the mathematics and science subareas separately in the order of difficulty or proportion correct. We then created a Form A that consists of the questions ranked 1, 6, 7, 12, 13, 18, 19, 24, 25, 30 in mathematics and the questions ranked 1, 6, 7, 12, 13, 18, 19, 24, 25, 30 in science. Thus, Form A has a total of 10 mathematics items and 10 science items. Note that it is not uncommon to find operationally reported subscores that are based on 10 or fewer items—see Figure 20 of Goodman and Hambleton (2004) for an example. Hence the choice of 10 items contributing to each subscore is quite realistic. Similarly, we created a Form B with questions ranked 2, 5, 8, 11, 14, 17, 20, 23, 26, 29 in mathematics and in science, and a Form C with the remaining questions. Forms A, B, and C can be considered roughly parallel forms and, by construction, all of the several thousand examinees who took the total test took all three of these forms. The subscore reliabilities on Forms A, B, and C range between .46 and .60. The subscore means on Form A, B, and C are 6.5 for mathematics and 5.7 for science.

We found that 6,035 examinees scored 4 (1st quartile) or lower on science on Form A. Such examinees will most likely be thought to be weak in science and provided additional instructions in science. Is this justified? We examined the science subscores of these 6,035 examinees on Forms B and C. Of them, 34% scored higher than the first quartile on science on Form B and 49% scored higher than the first quartile on science on Form C. We also found that of those who scored less than 5 (first quartile) or lower on mathematics on Form A, 30% scored higher than the first quartile on mathematics on Form B and 35% scored higher than the first quartile on mathematics on Form C.

We then identified all the 384 examinees who obtained a subscore of 8 on mathematics (third quartile) and 4 on science (first quartile) on Form A. Such examinees will most likely be thought to be strong in mathematics and weak in science and given additional instructions in science. We examined the mathematics and science subscores of these 384 examinees on Forms B and C. Table 1 shows a cross-tabulation of the subscores on Form B of these examinees. The table shows that researchers could often reach different conclusions based on the subscores on Form A and Form B. The percentage of examinees whose mathematics score is higher than their science score is only 68 on both Forms B and C. Also, of the 373 examinees who scored 5 (first quartile) on mathematics and 7 (third quartile) on science on form A, the percentage of examinees whose science score is higher than their mathematics score is 49 and 50, respectively, on Forms B and C.

This simple example demonstrates that remedial and instructional decisions based on subscores based on few items will have a high chance to be incorrect.

Reliability is defined as the correlation between the scores on a test and those on a parallel test and hence it is also possible to demonstrate the lack of usefulness of the 10-item subscores by computing simple correlations. The correlation of the science subscore on Form A (10 items) and the science subscore on Form B (10 items) is .42 while the correlation of the science subscore on Form A (10 items) and the total score on Form B (20 items) is .50—thus the total score on Form B is a much better predictor than the science subscore on Form B of the science subscore on the parallel Form A.

Alternatives to Simple Subscores

Researchers suggested several alternatives to simple subscores. These alternatives are briefly described next.

Augmented Subscores and Weighted Averages

Wainer et al. (2000) suggested an approach to increase the precision of a subscore by borrowing information from other

Table 1. Cross-Tabulations on Form B of the 384 Examinees with Mathematics Subscore of 8 and Science Subscore of 4 on Form A

	Science Subscore										
Math Subscore	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Total**
3	0	1	2	0	3	4	3	0	0	0	13
4	1	1	1	6	7	3	0	1	0	0	20
5	0	1	4	8	13	13	3	1	1	0	44
6	0	2	6	6	10	15	6	5	0	1	51
7	0	4	2	12	17	17	16	2	3	0	73
8	0	1	5	12	17	25	11	10	3	0	84
9	0	2	4	6	17	14	13	6	1	0	63
10	0	0	2	3	6	10	7	4	4	0	36
Total	1	12	26	53	90	101	59	29	12	1	384

subscores. Because subscores are almost always found to correlate moderately or highly with each other, it is reasonable to assume that, for example, the science subscore of a student provides some information about the math subscore of the same student. In this approach, an examinee's "augmented" subscore on a particular subscale (e.g., math) would be a function of that examinee's score on that subscale and that examinee's score on the remaining subscales (e.g., science, reading, etc). The subscales that have the strongest correlation with the math subscale have larger weights and have more influence on the "augmented" math subscore. Haberman (2008a) suggested the use a "weighted average" of the subscore and the total score to estimate the true subscore. This weighted average is a special case of the augmented subscore (Wainer et al., 2000). Based on notation introduced earlier, the weighted average can be denoted as

$$s_{sx} = \bar{s} + a(s - \bar{s}) + b(x - \bar{x}),$$

where a and b are constants that depend on the reliabilities and standard deviations of the subscore and the total score and the correlations between the subscores. For example, for the aforementioned form of the Elementary Education: Content Knowledge test, the values of a and b are .62 and .11, respectively. Hence, the weighted average corresponding to Mary D. Poppins's mathematics subscore, which is an estimate of her true mathematics subscore based on her observed mathematics subscore and her observed total score, is

$$19.4 + 0.62(12 - 19.4) + 0.11(75 - 78.4) = 14.42,$$

which is in between the estimate based on her observed subscore and the estimate based on her observed total score. The solid squares in Figure 2 show the weighted averages of the same nine examinees considered earlier for the Elementary Education: Content Knowledge test. Note that whenever an examinee obtained a subscore that is lower/higher compared to his/her total score, the corresponding weighted average differs considerably from the subscore (e.g., the mathematics subscore of those with total scores 75 and 91 and the social studies subscore of those with total score 44, 68, and 91).[1]

Haberman (2008a) and Haberman et al. (2009) discussed the computation of the PRMSE of augmented subscores and weighted averages. If the PRMSE of an augmented subscore or a weighted average is substantially larger than both the PRMSE_s and PRMSE_x of the corresponding subscore, then the augmented subscore or the weighted average can be reported. Whether an improvement in reliability (i.e., higher PRMSEs) for augmented subscores and weighted averages is large enough to justify reporting them instead of raw subscores depends on the purpose of the subscores. If the subscores are used for high stakes decisions such as licensing, then even a small gain in reliability (e.g., .01) may trigger the need to report augmented subscores or weighted averages instead of raw subscores. But in other cases (e.g., using subscores for identifying individual strengths and weaknesses on subscores), a larger gain in reliability may be required. For the aforementioned form of the Elementary Education: Content Knowledge test, the values of the PRMSE for the weighted averages are .80, .87, .79, and .83, respectively and the values of the PRMSE for the augmented subscores are .80, .87, .80, and .83, respectively. The PRMSE for any augmented subscore or weighted average is substantially larger than the corresponding PRMSE_s and PRMSE_x (see an earlier section for values of PRMSE_s and PRMSE_x for these data). Therefore, the augmented subscores or the weighted averages can be reported for the test.

It has been shown (e.g., Wainer et al., 2000, 2001; Puhan et al., 2010; Sinharay, 2010; Skorupski and Carvajal, 2010) that augmented subscores and weighted averages often are substantially more reliable than simple subscores and often have added value over simple subscores and the total score. Sinharay found that there is hardly any difference between the augmented subscore and the weighted average. Researchers such as Stone et al. (2010) and Skorupski and Carvajal claimed that there are problems with the validity of augmented subscores, but Sinharay, Haberman, and Wainer (2011) argued that no such problems existed.

Objective Performance Index

The objective performance index (Yen, 1987) is another approach to enhancing a subscore by borrowing information from other parts of the test. This approach uses a combination of IRT and Bayesian methodology. To compute the objective performance index, one first applies a unidimensional IRT model to compute an estimate of the subscore of each examinee, which is expressed as a percent-correct score, based on the examinee's overall test performance. A χ^2-type statistic is used to determine if the estimated subscore differs significantly from the observed subscore, which is also expressed as a percent-correct score. If the estimated subscore does not differ significantly from the observed subscore, then the objective performance index is defined as a weighted average of the observed subscore and the estimated subscore. If, however, the estimated subscore differs significantly from the observed subscore, then the objective performance index is defined as the observed subscore. See Figure 19 of Goodman and Hambleton (2004) for an example of an operational score report that includes the objective performance indices of an examinee. It should be noted that this approach, because of the use of a unidimensional IRT model, may not provide accurate results when the data are truly multidimensional. Ironically, that is when subscores can be expected to have added value.

Estimates Skill Parameters from a Cognitive Diagnostic Model

It is possible to employ a psychometric model such as a cognitive diagnostic model (e.g., Fu & Li, 2007) or a diagnostic classification model (Rupp & Templin, 2008) to report diagnostic scores instead of reporting subscores. See, for example, DiBello, Roussos, and Stout (2007), Leighton and Gierl (2007), Rupp, Templin, and Henson (2010), and von Davier, DiBello, and Yamamoto (2008) for further details on these models. These models assume that (a) solving each test item requires one or more skills, (b) each examinee has a discrete latent skill parameter corresponding to each of the skills, and (c) the probability that an examinee will answer an item correctly is a mathematical function of the skills the item requires and the latent skill parameters of the examinee. After a diagnostic classification model is fitted to a data set, the estimated values of the skill parameters are the diagnostic scores that can be reported. Examples of such models are the rule space model (RSM; Tatsuoka, 1983), the attribute hierarchy method (AHM; Leighton, Gierl, & Hunka, 2004), the DINA and NIDA models (Junker & Sijtsma, 2001), the general diagnostic model (von Davier, 2008), and the reparameterized unified model (Roussos et al., 2007). The first two of these, the RSM and AHM, are slightly different from the other diagnostic classification models in nature because they do not estimate any skill parameters—they match the response pattern of each examinee to several ideal or expected response patterns to determine what skills the examinee possesses. While there has been substantial research on diagnostic classification models, as Rupp and Templin (2009) acknowledge, there has not been a very convincing case that unequivocally illustrates how the added parametric complexity of these models, compared to simpler measurement models, can be justified in practice. In addition, there have been few empirical illustrations that the diagnostic scores produced by these models are reliable and valid (see, e.g., Haberman & von Davier, 2007; Sinharay & Haberman, 2008b).

Estimates from a MIRT Model

Researchers, such as Luecht (2003), de la Torre and Patz (2005), and Haberman and Sinharay (2010) examined the reporting of estimated ability parameters using MIRT models (e.g., Reckase, 1997; Ackerman et al., 2003). Haberman and Sinharay suggested reporting of estimated true subscores, that are in the same scale as the number-correct subscores, using MIRT models. de la Torre and Patz and Haberman and Sinharay found that there is not much difference between MIRT-based diagnostic scores and augmented subscores (Wainer et al., 2001). Haberman and Sinharay also suggested a method to determine if subscores based on MIRT models are of added value.

Notes on Using a Psychometric Model to Report Diagnostic Scores

If a psychometric model is employed to report diagnostic scores rather than simply reporting subscores, we think that the burden of proof lies on the person applying the model to demonstrate that the model parameters can be reliably estimated, that the model approximates the observed pattern of responses better than a simpler model (e.g., a univariate IRT model), that the computations required to fit the model are not excessively time-consuming, and that the diagnostic scores reported by the model have added value over a simple subscore or over the score(s) reported by a simple model. The simplest of the models passing fulfilling the demands may be operationally used for diagnostic score reporting.

Comparison of Different Approaches

Researchers have also compared subscores and their alternatives in terms of accuracy in estimating the true subscores. For example, Dwyer, Boughton, Yao, Steffen, and Lewis (2006) compared raw subscores with three alternatives—objective performance index, augmented subscores, and MIRT-based subscores. They found that the MIRT-based and augmentation methods performed best overall in estimating the true subscore. Fu and Qu (2010) also found the MIRT-based and augmentation methods to be the best among several methods of reporting subscores they studied.

Recommendations

This module discussed different issues involving subscores. The following list provides our recommendations for those interested in reporting subscores:

- If you are building a test for which subscores are to be reported, use a technique such as evidence-centered design (e.g., Mislevy, Steinberg, & Almond, 2003) or assessment engineering practices for item and test design (e.g., Luecht, Gierl, Tan, & Huff, 2006) to ensure that the test would allow you to report subscores. Test developers will have an important role in this. For example, Wainer et al. (2000) found that the data from a test were unidimensional and commented that "This has an important message for test developers, specifically, that if they want to measure multiple orthogonal dimensions they need to profoundly modify their item-writing and test-construction techniques."
- Whenever subscores are provided, provide evidence of adequate reliability, validity, and distinctness of the subscores. Any reported subscore, in order to be reliable, should be based on a sufficient number of items. Combining some subscores may result in subscores that have higher reliability and hence added value (although it might make the definition of the subscore broader). For example, subscores for "Physics: theory" and "Physics: applications" may be combined to yield one subscore for "Physics." It is important in the planning stage to ensure that the skills of interest are as distinct as possible from each other (though this is quite a difficult task before seeing the data).
- Consider reporting weighted averages or augmented subscores that often have added value (e.g., Sinharay, 2010) and often provide more accurate diagnostic information than the subscores do. Weighted averages may be difficult to explain to the general public, who may not like the idea that, for example, a reported reading subscore is based not

only on the observed reading subscore, but also on the observed writing subscore. Several approaches to the issue of explaining such weighted averages can be considered. One is that the weighted average better estimates examinee proficiency in the content domain represented by the subscore than does the subscore itself. This result can be discussed in terms of prediction of performance on an alternative test. The issue can also be discussed in terms of common cases in which information is customarily combined. For example, premiums for automobile insurance reflect not just the driving experience of the policy holder but also related information (such as education and marital status) that predicts future driving performance. In most cases, this difficulty in explanation of the weighted averages is more than compensated for by the higher PRMSE (i.e., more precision) of the weighted average.

- This module has primarily centered on subscores for individual examinees. Subscores can be reported at an aggregate level as well. For example, several testing programs report average subscores for the institutions that the examinees belong to. See Sinharay, Puhan, and Haberman (2010) for an example of such a score report. From our experience, there seems to be a common misconception that if one computes an average (over examinees) of subscores that are based on only a few items, the errors will cancel out and the average subscore will be worth reporting. However, Longford (1990) and Haberman, Sinharay, and Puhan (2009) suggested methods to examine whether aggregate-level subscores are of added value, and presented examples of situations when aggregate-level subscores do not have added value. So, examine the quality of the aggregate-level subscores before reporting them.
- Report subscores on an established scale. A temptation may exist to make this scale comparable to the scale for the total score or equal to a fraction of the scale that corresponds to the relative importance of the subscore, but these choices are not without difficulties given that subscores and total scores typically differ in reliability. In addition, equate the reported subscores so that the definition of strong performance in a subject area does not change across different administrations of a test. In typical cases, equating is feasible for the total score but not for subscores (e.g., if an anchor test is used to equate the total test, only a few of the items will correspond to a particular subarea so that an anchor test equating of the corresponding subscore is not feasible). Some work on equating of subscores and weighted averages has been done by Puhan and Liang (2011) and Sinharay and Haberman (2011).
- It is possible to perform an analysis of residuals to find the examinees who scored much lower or higher in one subarea compared to the other subareas. For example, Haberman (2008a) considered a residual that is the difference between the true subscore and the linear regression of the true subscore on the true total score. However, as Haberman found, the reliability of residuals for existing educational tests is expected to be low. Therefore, it is not straightforward to find residuals that are trustworthy.
- Finally, remember the advice of Luecht et al. (2006) that "inherently unidimensional item and test information cannot be decomposed to produce useful multidimensional score profiles—no matter how well intentioned or which psychometric model is used to extract the information" and that we should not "try to extract something that is not there" (p. 6). Thus, for some tests, changing the structure, by using, for example, sound assessment engineering practices for item and test design (Luecht et al., 2006) may be the only option in order to be able to report subscores, or, more generally, any kind of diagnostic scores. If restructuring the test is not a reasonable option, then, instead of subscore reporting, one can consider alternatives such as scale anchoring (e.g., Beaton & Allen, 1992), which makes claims about what students at different score points know and can do, and item mapping (e.g., Zwick, Senturk, Wang, & Loomis, 2001), that involves the use of exemplar items to characterize particular score points.

Self-Test[2]

1. Proportional Reduction of Mean Squared Errors: Theory

Let us denote the observed total score and the true total score of an examinee by x and x_t, respectively. Let us denote the observed subscore and the true subscore of the examinee by s and s_t. Any reported subscore for the examinee can be viewed as an estimate of the true subscore. Let us denote any such estimate by e. The mean squared error of e is given by $E(e - s_t)^2$. A baseline estimate is the trivial estimate $E(s)$.

(a) Prove that if e is restricted to a constant, the mean square is minimized when the constant is equal to $E(s)$.
The proportional reduction of mean squared error (PRMSE) due to the use of e instead of the use of $E(s)$ is given by

$$\frac{E(E(s) - s_t)^2 - E(e - s_t)^2}{E(E(s) - s_t)^2}.$$

Now let us consider two estimates of the true subscore:

- s_s, the regression of the true subscore s_t on the observed subscore s.
- s_x, the regression of the true subscore s_t on the total score x.

(b) Derive a simple form for s_s.
(c) Prove that the PRMSE of s_s is identical to the reliability of the subscore.
(d) Compute the PRMSE of s_x.

If the subscore has added value over the total score, then s_s should perform better than s_x in predicting s_t. In other words, the MSE for s_s should be smaller than s_x, or, the PRMSE of s_s should be larger than s_x.

2. Proportional Reduction of Mean Squared Errors: Application

Consider the two subscores considered in Ackerman and Shu (2009) for a fifth grade end-of-grade assessment. The reliabilities of the subscores (one measuring ability to understand the meaning of words and phrases, with 55 items, and the other measuring comprehension with 18 items) are .85 and .59, respectively. The standard deviation of the subscores are 6.65 and 2.05, respectively, the correlation between the subscores is .65 and the total test reliability is .87. Compute the PRMSE for s_x and comment if the subscores have added value for the data set.

3. Equating Subscores: Additional Issues

A testing practitioner is thinking of ways to equate (a) raw subscores across parallel forms of the same test, and (b) augmented subscores across parallel forms of the test. A colleague suggested the use of the equated total scores (which are comparable across the new and old forms) as an anchor score to equate the subscores. Which of the two equating procedures (i.e., equating raw subscores or augmented subscores) will benefit more from using the total equated score as an anchor score?

4. True/False Questions

(a) Subscores are most likely to have added value if they have relatively high reliability and if the true subscore and true total score have only a moderate correlation.

(b) A unidimensional test cannot provide useful multidimensional score profiles, no matter which psychometric model is used.

(c) I have a test for which the subscores are neither reliable nor distinct—so I understand that the subscores should not be reported for individual examinees. However, I can report the averages of these subscores for universities that the examinees belong to because when I average, the errors would cancel out.

(d) I have a test in which the subscores are based on a few items and do not have added value. However, many people who use these subscores say that they help examinees find their weaknesses—so I believe that the subscores should be reported.

(e) The correlation coefficient between two subscores is .45—so they must be distinct.

(f) Equating subscores guarantees their usefulness for diagnostic purposes.

Answers

1.

(a) Suppose e is a constant

$$\begin{aligned}
E(e-s_t)^2 &= E[e-E(s)+E(s)-s_t]^2\\
&= E[e-E(s)]^2+E[E(s)-s_t]^2\\
&\quad +2E\{[e-E(s)][E(s)-s_t]\}\\
&= E[e-E(s)]^2+E[E(s)-s_t]^2\\
&\quad +2[e-E(s)]E[E(s)-s_t]\\
&\quad \text{because } e \text{ is a constant}\\
&= E[e-E(s)]^2+E[E(s)-s_t]^2\\
&\quad +2[e-E(s)][E(s)-E(s_t)]\\
&= E[e-E(s)]^2+E[E(s)-s_t]^2\\
&\quad +2[e-E(s)][E(s)-E(s)]\\
&= E[e-E(s)]^2+E[E(s)-s_t]^2\\
&\ge E[E(s)-s_t]^2,\\
&\quad \text{where equality holds if and only if } e=E(s).
\end{aligned}$$

(b) The linear regression of v on u is given by $E(v|u) = E(v) + \rho(u,v)\frac{\sqrt{\mathrm{Var}(v)}}{\sqrt{\mathrm{Var}(u)}}[u - E(u)]$, where $\rho(u,v)$ denotes correlation between u and v. Therefore, s$_s$, the linear regression of s_t on s is given by $s_s = E(s_t) + \rho(s_t,s)\frac{\sqrt{\mathrm{Var}(s_t)}}{\sqrt{\mathrm{Var}(s)}}[s-E(s)] = E(s)+r_s[s-E(s)]$, where r_s is the reliability of the subscore, because $E(s_t) = E(s)$ and $\rho(s_t,s) = \frac{\sqrt{\mathrm{Var}(s_t)}}{\sqrt{\mathrm{Var}(s)}} = \sqrt{r_s}$.

(c) The PRMSE of s_s is given by

$$\begin{aligned}
&\frac{E(E(s)-s_t)^2-E(s_s-s_t)^2}{E(E(s)-s_t)^2}\\
&\quad=\frac{\mathrm{Var}(s_t)-E\{E(s_t|s)-s_t\}^2}{\mathrm{Var}(s_t)}\\
&\quad=\frac{\mathrm{Var}(s_t)-\mathrm{Var}(s_t|s)}{\mathrm{Var}(s_t)}\\
&\quad=\frac{\mathrm{Var}(s_t)-\mathrm{Var}(s_t)[1-\rho^2(s_t,s)]}{\mathrm{Var}(s_t)}=\rho^2(s_t,s)=r_s.
\end{aligned}$$

(d) The PRMSE of s_x is given by

$$\begin{aligned}
&\frac{E(E(s)-s_t)^2-E(s_x-s_t)^2}{E(E(s)-s_t)^2}\\
&\quad=\frac{\mathrm{Var}(s_t)-E\{E(s_t|x)-s_t\}^2}{\mathrm{Var}(s_t)}\\
&\quad=\frac{\mathrm{Var}(s_t)-\mathrm{Var}(s_t|x)}{\mathrm{Var}(s_t)}\\
&\quad=\frac{\mathrm{Var}(s_t)-\mathrm{Var}(s_t)[1-\rho^2(s_t,x)]}{\mathrm{Var}(s_t)}\\
&\quad=\rho^2(s_t,x)=\frac{[\mathrm{Cov}(s_t,x)]^2}{\mathrm{Var}(s_t)\mathrm{Var}(x)}\\
&\quad=\frac{[\mathrm{Cov}(s_t,x_t)]^2}{\mathrm{Var}(s_t)\mathrm{Var}(x)}\\
&\quad=\frac{[\mathrm{Cov}(s_t,x_t)]^2}{\mathrm{Var}(s_t)\mathrm{Var}(x_t)}\frac{\mathrm{Var}(x_t)}{\mathrm{Var}(x)}=\rho^2(s_t,x_t)\rho^2(x_t,x).
\end{aligned}$$

If the subscore has added value over the total score, then r$_s$ should be larger than $\rho^2(s_t,x_t)\rho^2(x_t,x)$. See, for example, Sinharay and Haberman (2008a), for details on the computation of $\rho^2(s_t,x_t)$.

2.

Var(x) = Sum of the variances of the subscores+2×correlation between the subscores × product of the standard deviations of the subscores = 66.15.

Var(x_t) = Var(x)×Total score reliability = 57.55.

Variance between the true subscores =

$$\begin{pmatrix} 6.65^2\times 0.85 & 0.65\times 6.65\times 2.05\\ 0.65\times 6.65\times 2.05 & 2.05^2\times 0.59 \end{pmatrix}=\begin{pmatrix} 37.59 & 8.86\\ 8.86 & 2.48\end{pmatrix},$$

because (a) the covariance between two different true subscores is the same as the covariance between the corresponding observed subscores and (b) the variance of a true subscore is the variance of the corresponding observed subscore multiplied the reliability of the subscore.

$\mathrm{Cov}(s_t, x_t)$ for subscores = the sum of the corresponding row of the above variance matrix, and are equal to 46.45 and 11.34 for the two subscores.

$\rho^2(x_t, x) =$ Total score reliability = .87.

$$\rho^2(s_t, x_t) = \frac{[\mathrm{Cov}(s_t, x_t)]^2}{\mathrm{Var}(s_t)\mathrm{Var}(x_t)}$$

$$= \frac{(46.45)^2}{37.59 \times 57.55} \quad \text{and} \quad \frac{(11.34)^2}{2.48 \times 57.55} \text{ for the two subscores}$$

$$= .997 \text{ and } .901 \text{ for the two subscores.}$$

So, PRMSE for $s_x = \rho^2(s_t, x_t)\rho^2(x_t, x) = .87$ and .78 for the two subscores.

Because both of these values are larger than the corresponding reliability values, the subscores do not have added value for this data set.[3]

3.

The equating of augmented subscores will benefit more than equating of raw subscores if total equated scores were used as an anchor. Augmented subscores borrow information from either the total score or other subscores. Therefore, compared to the raw subscores, augmented subscores will correlate more with the total score, which will lead to a more precise equating in terms of reduced random equating error.

4.

True/False questions

(a) True. Both conditions are important to have added value.

(b) True. A psychometric model cannot extract something that is not there (See Luecht et al., 2006).

(c) False. One has to analyze the average subscores for the universities to examine if they have any added value over average total scores (as in Haberman et al., 2009; Longford, 1990).

(d) False. Unreliable subscores may produce accurate remedial information for some examinees, but they will provide incorrect remedial information for many other examinees.

(e) False. It may be the case that these subscores have extremely low reliability and hence very high disattenuated correlation.

(f) False. The subscores have to be reliable and distinct from one another before equating can provide any added diagnostic benefits.

Acknowledgments

We are grateful to Mark Gierl, Matthias von Davier, Daniel Eignor, James Carlson, Todd Rogers, Xiaowen Zhu, and Richard Feinberg for helpful comments on earlier versions of this module.

Notes

[1] Note that the mathematics subscore mean is much larger than the other subscore means—so the mathematics subscore is expected to be much larger than the other subscores for all examinees. The weighted average for the mathematics subscore will differ from the subscore itself only if the latter is much lower/higher (compared to the other subscores) than the mathematics subscore mean.

[2] The first two questions of the self-test are intended for the readers who are comfortable with mathematical derivations.

[3] Note that this is true even though as many as 55 items contribute to the first subscore.

References

Ackerman, T. A., Gierl, M. J., & Walker, C. M. (2003), Using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement: Issues and Practice*, *22*, 37–51.

Ackerman, T., & Shu, Z. (2009, April). *Using confirmatory MIRT modeling to provide diagnostic information in large scale assessment*. Paper presented at the meeting of the National Council of Measurement in Education, San Diego, CA.

American Educational Research Association (AERA), American Psychological Association, National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: AERA.

Beaton, A. E., & Allen, N. L. (1992). Interpreting scales through scale anchoring. *Journal of Educational Statistics*, *17*, 191–204.

de la Torre, J., & Patz, R. J. (2005). Making the most of what we have: A practical application of multidimensional IRT in test scoring. *Journal of Educational and Behavioral Statistics*, *30*, 295–311.

DiBello, L. V., Roussos, L., & Stout, W. F. (2007). Review of cognitive diagnostic assessment and a summary of psychometric models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics*, *Volume 26* (pp. 977–1030). Amsterdam: Elsevier Science B.V.

Dwyer, A., Boughton, K. A., Yao, L., Steffen, M., & Lewis, D. (2006, April). *A comparison of subscale score augmentation methods using empirical data*. Paper presented at the meeting of the National Council on Measurement in Education, San Fransisco, CA.

Educational Tesing Service (2008). *Praxis*$^{\mathrm{TM}}$ *2008–09 information bulletin*. Princeton, NJ: Educational Testing Service.

Fu, J., & Li, Y. (2007, April). *Cognitively diagnostic psychometric models: An integrative review*. Paper presented at the meeting of the National Council on Measurement in Education, Chicago, IL.

Fu, J., & Qu, Y. (2010, April). *A comparison of subscore reporting approaches on simulated data*. Paper presented at the meeting of the National Council on Measurement in Education, Denver, CO.

Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. In I. Mervielde, I. Deary, F. De Fruyt, & F. Ostendorf (Eds.), *Personality psychology in Europe* (Vol. 7, pp. 7–28). Tilburg, The Netherlands: Tilburg University Press.

Goodman, D. P., & Hambleton, R. K. (2004). Student test score reports and interpretive guides: Review of current practices and suggestions for future research. *Applied Measurement in Education*, *17*, 145–220.

Haberman, S. J. (2008a). When can subscores have value? *Journal of Educational and Behavioral Statistics*, *33*, 204–229.

Haberman, S. J. (2008b). *Subscores and validity* (Research Report RR-08–64). Princeton, NJ: Educational Testing Service.

Haberman, S. J., & Sinharay, S. (2010). Reporting of subscores using multidimensional item response theory. *Psychometrika*, *75*, 209–227.

Haberman, S. J., Sinharay, S., & Puhan, G. (2009). Reporting subscores for institutions. *British Journal of Mathematical and Statistical Psychology*, *62*, 79–95.

Haberman, S. J., & von Davier, M. (2007). Some notes on models for cognitively based skills diagnosis. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics 26* (pp. 1031–1038). Amsterdam: Elsevier North-Holland.

Harris, D. J., & Hanson, B. A. (1991, March). *Methods of examining the usefulness of subscores*. Paper presented at the meeting of the National Council on Measurement in Education, Chicago, IL.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25*, 258–272.

Katz, I. R., Attali, Y., Rijmen, F., & Williamson, D. M. (2008, April). *ETS's iSkills$^{\text{TM}}$ assessment: Measurement of information and communication technology literacy*. Paper presented at the conference of the Society for Industrial and Organizational Psychology, San Francisco, CA.

Kim, H. R. (1994). *New techniques for the dimensionality assessment of standardized test data*. Unpublished doctoral dissertation, University of Illinois, Urbana-Champaign.

Leighton, J. P., & Gierl, M. J. (2007). *Cognitive diagnostic assessment for education: Theory and applications*. New York: Cambridge University Press.

Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy model for cognitive assessment: A variation on Tatsuoka's rule-space approach. *Journal of Educational Measurement*, *41*, 205–237.

Longford, N. T. (1990). Multivariate variance component analysis: An application in test development. *Journal of Educational Statistics*, *15*, 91–112.

Lord, F. M. (1965). A strong true-score theory with applications. *Psychometrika*, *30*, 239–270.

Luecht, R. M. (2003, April). *Applications of multidimensional diagnostic scoring for certification and licensure tests*. Paper presented at the meeting of the National Council on Measurement in Education, Chicago, IL.

Luecht, R. M., Gierl, M. J., Tan, X., & Huff, K. (2006, April). *Scalability and the development of useful diagnostic scales*. Paper presented at the annual Meeting of the National Council on Measurement in Education, San Francisco, CA.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, *1*, 3–67.

National Research Council (2001). *Knowing what students know: The science and design of educational assessment.* Washington, DC: The National Academies Press.

Puhan, G., & Liang, L. (2011). Equating subscores under the non equivalent anchor test (NEAT) design. *Educational Measurement: Issues and Practice*, *30*(1), 23–35.

Puhan, G., Sinharay, S., Haberman, S. J., Larkin, K. (2010). Comparison of subscores based on classical test theory. *Applied Measurement in Education*, *23*, 1–20.

Reckase, M. D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement*, *21*, 25–36.

Roussos, L. A., DiBello, L. V., Stout, W. F., Hartz, S. M., Henson, R. A., & Templin, J. H. (2007). The fusion model skills diagnostic system. In J. Leighton & M. Gierl (Ed.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 275–318). New York: Cambridge University Press.

Rupp, A. A., & Templin, J. L. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement*, *6*, 219–262.

Rupp, A. A., & Templin, J. L. (2009). The (un)usual suspects? A measurement community in search of its identity. *Measurement*, *7*(2), 115–121.

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York: Guilford Press.

Sinharay, S. (2010). How often do subscores have added value? Results from operational and simulated data. *Journal of Educational Measurement*, *47*, 150–174.

Sinharay, S., & Haberman, S. J. (2008a). *Reporting subscores: A survey* (Research Memorandum RM-08–18). Princeton, NJ: Educational Testing Service.

Sinharay, S., & Haberman, S. J. (2008b). How much can we reliably know about what students know? *Measurement: Interdisciplinary Research and Perspectives*, *6*, 46–49.

Sinharay, S., & Haberman, S. J. (2011). Equating of augmented subscores. *Journal of Educational Measurement*, *48*, 122–145.

Sinharay, S., Haberman, S. J., & Puhan, G. (2007). Subscores based on classical test theory: To report or not to report. *Educational Measurement: Issues and Practice*, *26* (4), 21–28.

Sinharay, S., Haberman, S. J., & Wainer, H. (2011). Adjusted subscores lack validity? Don't blame the messenger. *Educational and Psychological Measurement*. doi: 10.1177/0013164410391782 (published 22 March).

Sinharay, S., Puhan, G., & Haberman, S. J. (2010). Reporting diagnostic subscores in educational testing: Temptations, pitfalls, and some solutions. *Multivariate Behavioral Research*, *45*, 553–573.

Skorupski, W. P. & Carvajal, J. (2010). A comparison of approaches for improving the reliability of objective level scores. *Educational and Psychological Measurement*, 70, 357–375.

Stone, C. A., Ye, F., Zhu, X., & Lane, S. (2010). Providing subscale scores for diagnostic information: A case study when the test is essentially unidimensional. *Applied Measurement in Education*, *23*, 63–86.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, *52*, 589–617.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, *20*, 345–354.

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology*, *61*, 287–307.

von Davier, M., DiBello, L. V., & Yamamoto, K. (2008) Reporting test outcomes using models for cognitive diagnosis. In J. Hartig, E. Klieme, & D. Leutner (Eds.) *Assessment of competencies in educational contexts* (pp. 151–176). Cambridge, UK: Hogrefe & Huber.

Wainer, H., Sheehan, K. M., & Wang, X. (2000). Some paths toward making Praxis scores more useful. *Journal of Educational Measurement*, *37*, 113–140.

Wainer, H., Vevea, J. L., Camacho, F., Reeve, B. B., Rosa, K., Nelson, L., *et al*. (2001). Augmented scores—"borrowing strength" to compute scores based on small numbers of items. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 343–387). Mahwah, NJ: Erlbaum Associates.

Yen, W. M. (1987, June). *A Bayesian/IRT index of objective performance*. Paper presented at the meeting of the Psychometric Society, Montreal, Canada.

Zhang, J., & Stout, W. (1999). The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika*, *64*, 213–249.

Zwick, R., Senturk, D., Wang, J., & Loomis, S.C. (2001). An investigation of alternative methods for item mapping on the National Assessment of Educational Progress. *Educational Measurement: Issues and Practice*, *20*(2), 15–25.